Random Forest

Hoàng Vũ Minh

Random Forest là một thuật toán Machine Learning thuộc nhóm Supervised Learning, có thể được sử dụng cho cả bài toán hồi quy và phân loại.

Ý tưởng cơ bản:

Được gợi nhắc từ chính tên, Forest là một rừng cây, Random là ngẫu nhiên, Random Forest là một thuật toán theo phương pháp ensemble learning, bao gồm sự kết hợp kết quả từ nhiều Decision Tree thành kết quả đầu ra. Cho bài toán phân loại, có thể sử dụng voting. Nếu nhiều cây cho ra kết quả nào đó hơn, thì sẽ vote kết quả cuối cùng là nó. Với bài toán hồi quy, ta sử dụng trung bình.

Phương pháp lấy mẫu Boostrap

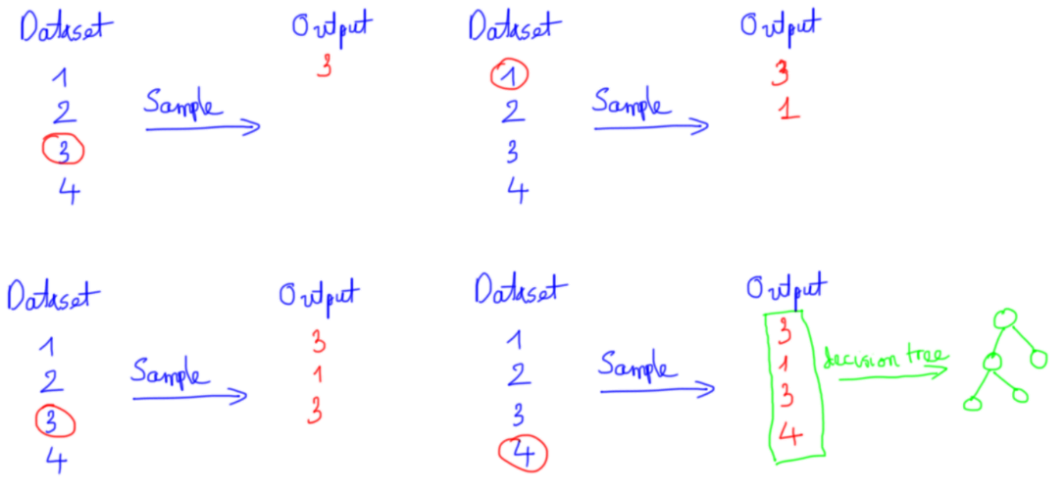
Để xây dựng được nhiều cây quyết định mà có ý nghĩa, ta cần sử dụng phương pháp nào đó nhằm lấy ra các mẫu khác nhau từ bộ dữ liệu gốc, và đó là boostrap. Phương pháp này còn gọi là Random Sampling with Replacement, có nghĩa là ta lấy mẫu đó ra xong lại bỏ nó lại vào tập ban đầu trước lần chọn tiếp theo. Bằng cách này, ta hoàn toàn có thể lấy được mẫu với những sample trùng lặp.

Ví dụ đơn giản: dataset ban đầu là D = {x1, x2, x3, x4, x5}

GIả sử ta cần lấy 1 mẫu boostrap 5 phần tử

- Lần 1: random trúng x1, mẫu là {x1}. Trả lại x1 vào phần chọn
- Lần 2: random trúng x2, mẫu là {x1, x2}
- Lần 3: random trúng x1, mẫu là {x1, x2, x1}
- Lần 4: random trúng x5, mẫu là {x1, x2, x1, x5}
- Lần 5: random trùng x, mẫu là {x1, x2, x1, x5, x1}

Vậy mẫu boostrap B là {x1, x2, x1, x5, x1}, sample x1 thậm chí được lặp lại 3 lần nhưng không có vấn dề gì. Sau đó ta xây dựng các cây quyết định dựa trên các mẫu này. Hình ảnh minh họa có thể được thể hiện qua hình dưới đây,



Giới thiệu Ensemble Learning:

Ensemble learning là một kỹ thuật quan trọng trong học máy. Nó cải thiện độ chính xác và tính đồng nhất bằng cách kết hợp các dự đoán từ nhiều mô hình khác nhau. Kỹ thuật này nhằm giảm thiểu sai số và khắc phục các thiên lệch có thể tồn tại trong từng mô hình riêng lẻ. Bằng cách sử dụng trí tuệ tổng hợp của toàn bộ ensemble (tập hợp) các mô hình, ensemble learning thường cho kết quả tốt hơn so với sử dụng một mô hình đơn lẻ và mang lại những dự báo hoặc dự đoán đáng tin cậy hơn.

Một vài kỹ thuật cơ bản nhất:

- Max Voting: Dùng trong bài toán phân loại, max voting xem xem các model dự đoán vào nhãn nào nhiều nhất, thì tổng hợp lại và đưa ra kết quả là nhãn đó.
- Averaging: Dùng trong bài toán hồi quy, mỗi model cho ra một số thì ta cộng lại lấy trung bình ra kqua cuối.
- Weighted Average: đấy là cách mở rộng của phàn trên, vẫn là lấy trung bình nhưng có trọng số để xác định xem kết quả của model nào ảnh hưởng nhất

NOTES: Đây chỉ là một vài cách cơ bản nhất, trên thực tế có rất nhiều cách Ensemble Learning phức tạp hơn rất nhiều, sẽ được trình bày ở phần về Deep Learning.

Nhận xét:

- Ưu điểm: có thể cho ra kết quả chính xác hơn Decision Tree, tránh overfitting. Cơ cơ hội tạo ra mô hình low bias, low variance.
- Kết quả đáng tin cậy hơn.